TIGER: A data analysis pipeline for testing the strong-field dynamics of general relativity with gravitational wave signals from coalescing compact binaries

M. Agathos[1], W. Del Pozzo[1,2], T.G.F. Li[1,3], C. Van Den Broeck[1], J. Veitch[1], S. Vitale[4]
[1] *Nikhef – National Institute for Subatomic Physics, Science Park 105, 1098 XG Amsterdam, The Netherlands*
[2] *School of Physics and Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom*
[3] *LIGO Laboratory, California Institute of Technology, Pasadena, CA 91125, USA*
[4] *LIGO Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
(Dated: November 5, 2013)

The direct detection of gravitational waves with upcoming second-generation gravitational wave detectors such as Advanced LIGO and Virgo will allow us to probe the genuinely strong-field dynamics of general relativity (GR) for the first time. We present a data analysis pipeline called TIGER (Test Infrastructure for GEneral Relativity), which is designed to utilize detections of compact binary coalescences to test GR in this regime. TIGER offers a model-independent test of GR itself, in that it is not necessary to compare with any specific alternative theory of gravity. It performs Bayesian inference on two hypotheses: the GR hypothesis $\mathcal{H}_{\rm GR}$, and $\mathcal{H}_{\rm modGR}$, which states that one or more of the post-Newtonian coefficients in the waveform are not as predicted by GR. By the use of multiple sub-hypotheses of $\mathcal{H}_{\rm modGR}$, in each of which a different number of parameterized deformations of the GR phase are allowed, an arbitrarily large number of "testing parameters" can be used without having to worry about a model being insufficiently parsimonious if the true number of extra parameters is in fact small. TIGER is well-suited to the regime where most sources have low signal-to-noise ratios, again through the use of these sub-hypotheses. Information from multiple sources can trivially be combined, leading to a stronger test. We focus on binary neutron star coalescences, since for such sources sufficiently accurate waveform models are available which can be generated fast enough on a computer that they can be used in Bayesian inference. By performing numerical experiments in Gaussian, stationary noise, we demonstrate that the pipeline is robust against a number of unknown fundamental, astrophysical, and instrumental effects, such as differences between waveform approximants, a limited number of post-Newtonian phase contributions being known, the effects of neutron star tidal deformability on the orbital motion, neutron star spins, and instrumental calibration errors.

PACS numbers: 04.80.Nn, 02.70.Uu, 02.70.Rr

I. INTRODUCTION AND OVERVIEW

General relativity (GR) is a highly non-linear, dynamical theory of gravity. Yet, until the 1970s, almost all of its tests were based on the behavior of test particles in a *static* gravitational field [1], such as the perihelion precession of Mercury, the deflection of starlight by the Sun, and Shapiro time delay. The parameterized post-Newtonian (PPN) formalism (for an overview, see [2]) was developed as a systematic framework for these and other tests; even so, the interpretation of most of the available data did not require much more than an expansion of the Schwarzschild metric in $GM/(c^2r)$, with M the mass and r the distance, up to the first few non-trivial orders. Although excellent agreement with theory was obtained, the tests that were actually performed amounted to little more than probing the effect on the motion of test masses of low-order general relativistic corrections to the Newtonian gravitational field.

The situation improved with the discovery of the Hulse-Taylor binary neutron star in 1974 [3]. One of the components could be observed electromagnetically as a pulsar, and this way it was inferred that the binary loses energy and angular momentum through gravitational wave (GW) emission as by GR, at least at the level of the quadrupole formula. Subsequently, more relativistic binaries were discovered, allowing for impressive new tests of GR in a parameterized post-Keplerian (PPK) framework [4]. However, if one is interested in further probing the dissipative dynamics of binaries, and especially the dynamics of spacetime itself, what matters is the *orbital compactness* $GM/(c^2R)$ (with M the total mass and R the separation), as well as the orbital velocity v/c. Even the newly discovered neutron star-white dwarf system [5] only has $GM/(c^2R) \sim 2\times10^{-6}$, and $v/c \sim 4\times10^{-3}$. For comparison, the surface gravity of the Sun is $GM_\odot/(c^2R_\odot) \sim 10^{-6}$, and the orbital velocity of Mercury is $v/c \sim 1.6\times10^{-4}$.

By contrast, binaries consisting of neutron stars and/or black holes on the verge of merger will have $GM/(c^2R) > 0.2$ and $v/c > 0.4$, with copious gravitational wave emission. Being able to observe the orbital motion of such systems would give us access to the genuinely strong-field, relativistic regime of gravity. Most importantly, we would like to probe the dynamical self-interaction of spacetime itself, such as the scattering of quadrupolar waves off the Schwarzschild curvature generated by the binary as a whole [6, 7]. The only way to gain empirical

access to such phenomena is through direct gravitational wave detection.

A network of second-generation gravitational wave detectors is currently under construction. The Advanced LIGO [8] and Advanced Virgo [9] GW observatories are expected to start taking data in 2015, with gradual upgrades in the following years. The smaller GEO-HF in Germany is already active [10]. KAGRA [11] in Japan and possibly LIGO-India [12] will come online a few years later. These detectors may find tens of GW signals per year from coalescing compact binaries composed of neutron stars and/or black holes. The predicted detection rates for the Advanced LIGO-Virgo network are in the range $1 - 100\,\mathrm{yr}^{-1}$ depending on the astrophysical event rate, the instruments' duty cycle, and the sensitivity evolution of the detectors [13, 14]; see also [15] for detection rate predictions assuming that short, hard gamma ray bursts are caused by coalescing binaries.

There is a considerable body of literature on the constraints that can be put on various *specific* alternative theories of gravity with ground-based and space-based GW detectors, and pulsar timing arrays; see [16, 17] and references therein. What we will be interested in here are *model-independent* tests of GR itself. A first step in that direction was taken by Arun *et al.* [18–20] in the context of compact binary inspiral. Their method exploits the fact that, at least for binaries where neither component has spin, all coefficients ψ_i in the post-Newtonian (PN) expansion of the inspiral phase (see below for their definition) only depend on the component masses m_1, m_2. Hence only two of them are independent, and a comparison of any three of them allows for a test of GR. Such a method would be extremely general, in that one does not have to look for any specific way in which GR might be violated; instead, very generic deviations can be searched for. A similar idea was pursued in the context of ringdown by Gossan *et al.* [21]: if the No Hair Theorem applies to Nature, then the frequencies f_{nlm} and damping times τ_{nlm} of the various ringdown modes again only depend on two quantities, in this case the mass M and spin J of the final black hole.

The original ideas of [18–20] have the drawback that they rely on parameter estimation, which makes it difficult to combine information from multiple sources. An alternative way of testing GR is *Bayesian model selection*. Here one compares two hypotheses, one corresponding to the GW waveform model predicted by GR, and the other to a model which has parameterized deformations of the GR waveform, characterized by additional parameters $\{q_1, q_2, \ldots, q_{N_T}\}$. This was the approach taken by Del Pozzo *et al.* [22] in the context of inspiral (where a single additional parameter was introduced, related to the graviton mass), and again by Gossan *et al.* for ringdown (where multiple extra free parameters were considered) [21]. Yunes and collaborators [23–25] proposed a parameterization of non-GR waveforms guided by the ways in which a variety of alternative theories of gravity modify the GR waveform, leading to the "parameterized post-Einsteinian" (PPE) framework. For the relationship between the PPN, PPK, and PPE formalisms, see [26].

In the abovementioned Bayesian studies, a comparison was made between a waveform model in which all the extra parameters q_i were allowed to vary, and a waveform model where all of them took their GR values (which for the present discussion we can take to mean $q_i = 0$ for $i = 1, \ldots, N_T$). As noted by Li *et al.* [27], this corresponds to asking the question "Do *all* of the q_i differ from zero at the same time?" Let us denote the associated hypothesis by $H_{12\ldots N_T}$, which is to be compared with the GR hypothesis $\mathcal{H}_{\rm GR}$. A more general (and hence more interesting) question is: "Do *one or more* of the q_i differ from zero?" Denote the corresponding hypothesis by $\mathcal{H}_{\rm modGR}$. As shown in [27], although there is no single waveform model associated with $\mathcal{H}_{\rm modGR}$, testing the latter amounts to testing $2^{N_T} - 1$ disjoint sub-hypotheses $H_{i_1 i_2 \ldots i_k}$ corresponding to all subsets $\{q_{i_1}, q_{i_2}, \ldots, q_{i_k}\}$ of the full set of "testing parameters" $\{q_1, q_2, \ldots, q_{N_T}\}$. A given $H_{i_1 i_2 \ldots i_k}$ is tested by a waveform model in which $q_{i_1}, q_{i_2}, \ldots, q_{i_k}$ are free, but all the other q_j are fixed to zero. The Bayes factors against GR for all of these sub-hypotheses can be combined into a single *odds ratio* which compares $\mathcal{H}_{\rm modGR}$ with $\mathcal{H}_{\rm GR}$.

As explained in [27–29] and further elucidated in this paper, the approach of Li *et al.* has several attractive features:

- One can use an arbitrarily large number of "testing parameters" without having to worry about a model being insufficiently parsimonious in cases where the true number of non-GR parameters is small, due to the availability of sub-hypotheses corresponding to different numbers of free parameters.
- Information from multiple sources can trivially be combined, leading to a stronger test of GR.
- It is well-suited to a regime where most sources have a small signal-to-noise ratio, again because of the use of multiple non-GR sub-hypotheses.
- It will allow us to find a wide range of deviations from GR, even ones that are well outside the particular parameterized waveform family used.
- The method is not tied to any given waveform model, or even any particular part of the coalescence process.

Given these advantages, it is natural to take the above scheme as a basis for computer code to test GR using actual detector data. Such a data analysis pipeline is now in place within the LIGO Algorithm Library [30]. It is called *TIGER*, for "Test Infrastructure for GEneral Relativity".

Before we can be sure of the usefulness of TIGER in a realistic data analysis setting, we must check its robustness against any unknown fundamental, astrophysical, and instrumental effects. We focus on BNS, since for

this case, waveform models that accurately capture the relevant physics *and* can be generated sufficiently fast on a computer have been available for some time now [31]. Eventually we would also like to extend TIGER to binary black hole (BBH) and possibly neutron star-black hole (NSBH) coalescences. For BBH, Pan *et al.* [32] recently presented a (semi-)analytic waveform model based on the Effective One-Body (EOB) formalism that covers the entire coalescence process (inspiral, merger, and ringdown) with all the relevant physics included (notably precessing spins) and which has been tuned to have an extremely high overlap with numerical relativity simulations. This will be a very useful signal model, but for "recovery" of signals, thousands of trial waveforms need to be compared with the data, which is computationally expensive. While the present paper was being finished, Hannam *et al.* [33] proposed a *frequency domain* inspiral-merger-ringdown waveform for BBH which captures precessing spins and can be generated very fast; this approximant, or a further improvement of it, may well be suitable as a recovery waveform for testing GR. Alternatively, recovery could be done with (an upgraded version of) the fast time domain "PhenSpin" approximant of Sturani *et al.* [34, 35]. However, for now we restrict the discussion to binary neutron star (BNS) coalescence, where the simpler post-Newtonian waveform approximants are already extremely accurate. This includes an analytic frequency domain model ("TaylorF2") whose calculation takes sufficiently little computational time.

Focusing on BNS, the following issues need to be addressed:

- Even for binary neutron star coalescence, there are small differences between the various post-Newtonian waveform approximants, as well between them and EOB models. Since TIGER is specifically designed to find anomalies in the signals, we must make sure that these discrepancies, however minor, are not mistaken for violations of GR.

- Post-Newtonian waveforms are only available up to 3.5PN in phase. What might be the effect of unknown PN contributions?

- In the final stages of inspiral, neutron stars get deformed because of each other's tidal fields. This has an effect on the orbital motion, which gets imprinted onto the GW signal waveform. The size of these tidal effects is set by the neutron star equation of state, about which currently not much is known. Can we avoid mistaking unknown tidal effects for a violation of GR?

- The dimensionless spins of neutron stars in binaries are generally expected to be quite small, but the resulting spin-orbit and spin-spin effects will nevertheless need to be taken into account.

- The calibration of the instruments will be imperfect, leading to frequency dependent uncertainties in the interpretation of amplitudes and phases. What will their impact be?

In order to see how these effects can be brought under control, we perform numerical experiments in simulated stationary, Gaussian noise following the predicted noise curves of Advanced LIGO and Advanced Virgo at their final design sensitivities [8, 9]. Note that in reality, the noise will be neither Gaussian nor stationary due to "glitches". As we will explain, TIGER involves the calculation of a so-called *background distribution*, in which these additional unknowns will be included automatically. However, here we focus on the points above; further instrumental issues will be dealt with in a forthcoming study.

This paper is structured as follows. In Sec. II, we outline the analysis pipeline. In the interest of having a self-contained discussion of TIGER as a whole, we first summarize the method of Li *et al.*, endeavoring to highlight the main features; an in-depth discussion can be found in [27, 28]. We explain how to calculate the main quantities of interest and how to deal with detector noise. Since the focus will be on BNS, we will need to explain how one would select for such sources in particular, making sure NSBH and BBH events are not mistakenly taken into account. Then the overall structure of the TIGER pipeline is discussed. We highlight the advantages of TIGER over more basic model selection methods, discuss its performance in the low signal-to-noise ratio regime, and give a sense of how generic it may be in finding violations of GR. The main results of this paper are presented in Sec. III, where we show how TIGER can be made robust against differences between waveform approximants, limited availability of post-Newtonian phase contributions, unknown neutron star tidal deformability, instrumental calibration errors, and the effects of neutron star spins. Conclusions and future directions are discussed in Sec. IV.

Unless stated otherwise, we will use units such that $G = c = 1$.

II. THE TIGER PIPELINE

A. Basic method

The core method of TIGER is that of Li *et al.*, which we briefly discuss; detailed derivations can be found in [27, 28].

Given a set of detected sources, we will want to use them to compare two hypotheses: $\mathcal{H}_{\rm GR}$, which says that the signal waveform is as predicted by GR, and $\mathcal{H}_{\rm modGR}$, which says that there is a deviation from GR. Ideally, $\mathcal{H}_{\rm modGR}$ would be the negation of $\mathcal{H}_{\rm GR}$, but evaluating such a hypothesis would involve checking the data against an infinite-dimensional waveform family corresponding to

the infinitely many ways in which the waveform might deviate from GR. Hence we need to consider a more limited hypothesis $\mathcal{H}_{\rm modGR}$, which will be based on the phasing. In the stationary phase approximation [36, 37], the gravitational wave phase takes the general form

$$\Psi(f) = 2\pi f t_c - \varphi_c - \frac{\pi}{4} + \sum_{j=0}^{7} \left[\psi_j + \psi_j^{(l)} \ln f\right] f^{(j-5)/3}, \tag{1}$$

where t_c and φ_c are, respectively, the time and phase at coalescence, and in GR, the coefficients ψ_j, $\psi_j^{(l)}$ are specific, known functions of the component masses m_1, m_2 and spins $\vec{S}_1$, $\vec{S}_2$. We now define hypotheses as follows:

- $\mathcal{H}_{\rm GR}$ is the hypothesis that the ψ_j, $\psi_j^{(l)}$ depend on masses and spins as predicted by GR.
- $\mathcal{H}_{\rm modGR}$ is the hypothesis that *one or more* of the ψ_j, $\psi_j^{(l)}$ deviate from the GR prediction, without specifying which.

We note that in principle, we could also have allowed for free parameters in the amplitude [23, 25]. However, with second-generation detectors and for stellar mass binaries, one will not have much sensitivity to subdominant amplitude effects [38–40].

There is no waveform model associated with $\mathcal{H}_{\rm modGR}$ as defined above. This is solved by introducing the following sub-hypotheses:

> $H_{i_1 i_2 \ldots i_k}$ is the hypothesis that the parameters $\{\psi_{i_1}, \psi_{i_2}, \ldots, \psi_{i_k}\}$ *do not* have the dependence on masses and spins as in GR, but all other coefficients $\psi_j \notin \{\psi_{i_1}, \psi_{i_2}, \ldots, \psi_{i_k}\}$ *do* depend on masses and spins as predicted by GR.

These hypotheses do have waveforms associated with them. Let $\vec{\theta}$ be the parameters that appear in the GR waveform (masses, spins, sky position, orientation, and distance). Then the hypothesis $H_{i_1 i_2 \ldots i_k}$ is tested by waveforms in which the parameters $\{\vec{\theta}, \psi_{i_1}, \psi_{i_2}, \ldots, \psi_{i_k}\}$ are allowed to vary freely, but the $\psi_j \notin \{\psi_{i_1}, \psi_{i_2}, \ldots, \psi_{i_k}\}$ are set to their GR expressions. It will be convenient to write the free parameters as

$$\psi_i = [1 + \delta\chi_i]\, \psi_i^{\rm GR}, \tag{2}$$

where $\psi_i^{\rm GR} = \psi_i^{\rm GR}(m_1, m_2, \vec{S}_1, \vec{S}_2)$ is the expression for ψ_i as a function of m_1, m_2, $\vec{S}_1$, $\vec{S}_2$ that GR predicts. Now, the hypothesis $\mathcal{H}_{\rm modGR}$ is the logical "or" of all the $H_{i_1 i_2 \ldots i_k}$:

$$\mathcal{H}_{\rm modGR} = \bigvee_{i_1 < i_2 < \ldots < i_k} H_{i_1 i_2 \ldots i_k}. \tag{3}$$

Given detections $d_1, d_2, \ldots, d_{\mathcal{N}}$, we define an *odds ratio*

$$\mathcal{O}^{\rm modGR}_{\rm GR} \equiv \frac{P(\mathcal{H}_{\rm modGR}|d_1, \ldots, d_{\mathcal{N}}, I)}{P(\mathcal{H}_{\rm GR}|d_1, \ldots, d_{\mathcal{N}}, I)}, \tag{4}$$

where $P(\mathcal{H}_{\rm modGR}|d_1, \ldots, d_{\mathcal{N}}, I)$ is the posterior probability for the hypothesis $\mathcal{H}_{\rm modGR}$ given the data $d_1, \ldots, d_{\mathcal{N}}$ and any background information I we may hold, and similarly for $P(\mathcal{H}_{\rm GR}|d_1, \ldots, d_{\mathcal{N}}, I)$. Using Eq. (3) and noting that the $H_{i_1 \ldots i_k}$ are all logically disjoint, assuming independence of the data streams $d_1, \ldots, d_{\mathcal{N}}$, and setting the prior odds $P(H_{i_1 \ldots i_k}|I)$ for all the sub-hypotheses equal to each other, repeated application of Bayes' theorem yields

$$\mathcal{O}^{\rm modGR}_{\rm GR} = \frac{\alpha}{2^{N_T} - 1} \sum_{i_1 < \ldots < i_k; k \le N_T} \prod_{A=1}^{\mathcal{N}} \frac{P(d_A|H_{i_1 \ldots i_k}, I)}{P(d_A|\mathcal{H}_{\rm GR}, I)}. \tag{5}$$

Here $\alpha = P(\mathcal{H}_{\rm modGR}|I)/P(\mathcal{H}_{\rm GR}|I)$, *i.e.* the ratio of prior odds for $\mathcal{H}_{\rm modGR}$ against $\mathcal{H}_{\rm GR}$; as we shall see, the choice of this overall prefactor will be irrelevant within the TIGER framework. N_T is the total number of coefficients that are allowed to vary freely. For reasons of computational expense, we will only allow the phase coefficients within a particular set $\{\psi_1, \psi_2, \ldots, \psi_{N_T}\}$ to differ from their GR predictions; these will be called our *testing parameters*.

From Eq. (5), we see that the quantities to be computed from the data are the *Bayes factors*

$${}^{(A)}B^{i_1 \ldots i_k}_{\rm GR} \equiv \frac{P(d_A|H_{i_1 \ldots i_k}, I)}{P(d_A|\mathcal{H}_{\rm GR}, I)}. \tag{6}$$

The *evidences* for the $H_{i_1 \ldots i_k}$ are given by

$$\begin{aligned} &P(d_A|H_{i_1 \ldots i_k}, I) \\ &= \int d\vec{\theta}\, d\delta\chi_{i_1} \ldots d\delta\chi_{i_k}\, \pi_{\rm GR}(\vec{\theta}|I)\, \pi_{i_1 \ldots i_k}(\delta\chi_{i_1}, \ldots, \delta\chi_{i_k}|I) \\ &\qquad p_{i_1 \ldots i_k}(d_A|\vec{\theta}, \delta\chi_{i_1}, \ldots, \delta\chi_{i_k}, I). \end{aligned} \tag{7}$$

Here $\pi_{\rm GR}(\vec{\theta}|I)$ is the prior density on the parameters in the GR waveform. We will let the prior on the chirp mass be of the form $p(\mathcal{M}|I) \propto \mathcal{M}^{-11/6}$ (see [41–43] for motivation) and the components masses m_1, m_2 are restricted to the interval $[1, 35]\, M_\odot$. The priors on sky position and orientation are taken to be uniform on the corresponding spheres. Distance is allowed to vary between 1 and 1000 Mpc, and the prior is uniform in volume. For the phase at coalescence, φ_c, we choose a flat prior on the interval $[0, 2\pi)$. Finally, the coalescence time t_c has a flat prior with a width of 100 ms around the value indicated by the search pipeline that made the detection. We let the prior densities $\pi_{i_1 \ldots i_k}(\delta\chi_{i_1}, \ldots, \delta\chi_{i_k}|I)$ on the $\delta\chi_i$ take the form

$$\pi_{i_1 \ldots i_k}(\delta\chi_{i_1}, \ldots, \delta\chi_{i_k}|I) = \prod_{a=1}^{k} \pi_{i_a}(\delta\chi_{i_a}|I). \tag{8}$$

For the purposes of this paper we choose all the $\pi_{i_a}(\delta\chi_{i_a}|I)$ to be equal to each other, and flat on the interval $[-0.5, 0.5]$, corresponding to the largest GR violations considered in our simulations. However, it should

be stressed that for use on real detections, one may want to take these intervals to be much wider. Indeed, existing observations offer no significant constraints on the phase parameters beyond 0PN, as even the electromagnetically observed binary neutron stars only probe the quadrupole formula. Next, consider the *likelihood functions* $p_{i_1\ldots i_k}(d_A|\vec{\theta},\delta\chi_{i_1},\ldots,\delta\chi_{i_k},I)$. For a single detector, these are given by

$$p_{i_1\ldots i_k}(d_A|\vec{\theta},\delta\chi_{i_1},\ldots,\delta\chi_{i_k},I)\propto\exp\left[-2\int_{f_0}^{f_{\rm LSO}}df\,\frac{|\tilde{d}_A(f)-\tilde{h}_{i_1\ldots i_k}(\vec{\theta},\delta\chi_{i_1},\ldots,\delta\chi_{i_k};f)|^2}{S_n(f)}\right], \tag{9}$$

where the proportionality factor is set by normalization, $\tilde{d}_A$ is the Fourier transform of the data stream, and $\tilde{h}_{i_1\ldots i_k}$ is the frequency domain waveform corresponding to the hypothesis $H_{i_1\ldots i_k}$. The detectors' lower cut-off frequency f_0 is taken to be 20 Hz, and $f_{\rm LSO}\equiv 1/(6^{3/2}\pi\,M)$, with M the total mass, is the frequency at last stable orbit in the limit where one of the components is a test particle. The one-sided noise power spectral density $S_n(f)$ can be estimated from the stretch of data immediately following the one containing the detection. Again up to normalization, the likelihood for a *network* of detectors is the product of the likelihoods for the individual interferometers, taking into account relative time shifts due to the detectors' different locations on the Earth. To probe the likelihood functions we use the nested sampling algorithm as implemented by Veitch and Vecchio [41–43], and we refer the reader to those papers for details.

The evidence for $\mathcal{H}_{\rm GR}$ is

$$P(d_A|\mathcal{H}_{\rm GR},I)=\int d\vec{\theta}\,\pi_{\rm GR}(\vec{\theta}|I)\,p_{\rm GR}(d_A|\vec{\theta},I). \tag{10}$$

The likelihood $p_{\rm GR}(d_A|\vec{\theta},I)$ is defined analogously to (9); it too is explored using nested sampling [41–43].

Finally, what our algorithm computes directly is not $B^{i_1\ldots i_k}_{\rm GR}$, but the Bayes factors $B^{i_1\ldots i_k}_{\rm noise}$ and $B^{\rm GR}_{\rm noise}$ for the hypotheses $H_{i_1\ldots i_k}$, $\mathcal{H}_{\rm GR}$ against the noise-only hypothesis $\mathcal{H}_{\rm noise}$, which states that there is no signal in the data. However, the desired quantities are obtained from the latter through $B^{i_1\ldots i_k}_{\rm GR}=B^{i_1\ldots i_k}_{\rm noise}/B^{\rm GR}_{\rm noise}$.

B. Dealing with noise

If GR happens to be valid then one would expect $\mathcal{O}^{\rm modGR}_{\rm GR}<1$, or $\ln\mathcal{O}^{\rm modGR}_{\rm GR}<0$. However, the noise in the detectors can mimic violations of GR, so that one can have $\ln\mathcal{O}^{\rm modGR}_{\rm GR}>0$ even if GR is in fact the correct theory of gravity. To make sure that we will not erroneously declare a GR violation, the measured log odds ratio will be compared with a *background distribution*. The latter is constructed by taking a large number of simulated GR signals, all having different masses, sky locations, orientations, and distances picked from astrophysically motivated distributions (see Sec. III below), and injecting them into stretches of data surrounding the ones the detections are in, to have similar noise realizations. Here one can adopt the treatment of "on-source" and "off-source" data as in searches for gravitational wave events associated with gamma ray bursts; see [44] and references therein. These injections can be combined randomly into catalogs of $\mathcal{N}$ sources each, with $\mathcal{N}$ the number of sources that were observed in reality. For each of these catalogs, one can compute $\ln\mathcal{O}^{\rm modGR}_{\rm GR}$, arriving at an estimate for the distribution of the log odds ratio for the case where GR is correct. Let us denote this distribution by $P(\ln\mathcal{O}|\mathcal{H}_{\rm GR},\kappa_{\rm GR},I)$, where $\kappa_{\rm GR}$ indicates the particular set of simulated signals, or "injections". Given such a distribution and picking a *maximum tolerable false alarm probability* β, a *threshold* $\ln\mathcal{O}_\beta$ can be computed for the measured log odds ratio to overcome. In the limit of infinitely many injections, $\ln\mathcal{O}_\beta$ is defined implicitly by

$$\beta=\int_{\ln\mathcal{O}_\beta}^{\infty}P(\ln\mathcal{O}|\mathcal{H}_{\rm GR},\kappa_{\rm GR},I)\,d\ln\mathcal{O}. \tag{11}$$

The background distribution and threshold for a given false alarm probability will also allow us to assess our ability to uncover a given type of GR violation. Let $\mathcal{H}_{\rm alt}$ denote the associated hypothesis. Then a *foreground distribution* of log odds ratio, $P(\ln\mathcal{O}|\mathcal{H}_{\rm alt},\kappa_{\rm alt},I)$, can be computed by analyzing a large number of (catalogs of) simulated signals, $\kappa_{\rm alt}$. The *efficiency* in finding the particular GR violation is defined as

$$\zeta=\int_{\ln\mathcal{O}_\beta}^{\infty}P(\ln\mathcal{O}|\mathcal{H}_{\rm alt},\kappa_{\rm alt},I)\,d\ln\mathcal{O}. \tag{12}$$

These definitions are illustrated in Fig. 1.

We can now see why the choice of prior odds $\alpha=P(\mathcal{H}_{\rm modGR}|I)/P(\mathcal{H}_{\rm GR}|I)$ entering Eq. (5) is irrelevant: A change $\alpha\to\alpha'$ simply causes a rigid shift by $\ln(\alpha'/\alpha)$ of the background distribution, the threshold, and the foreground distribution, which, however, leaves false alarm probabilities and efficiencies unaffected.

As we shall see in Sec. III B, for BNS the TaylorF2 approximant is sufficiently reliable as a recovery waveform.

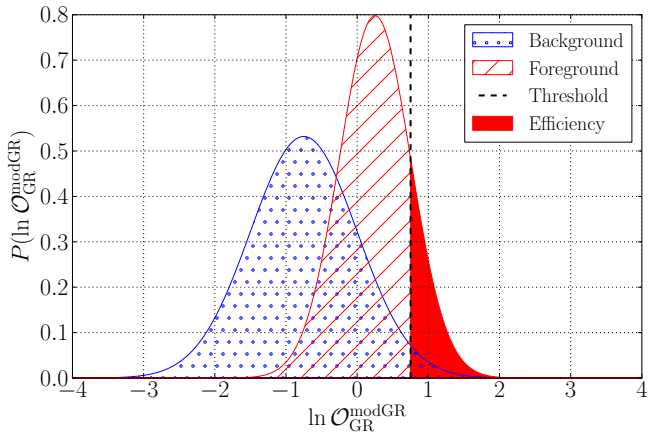

FIG. 1: A schematic illustration of background distribution, threshold, foreground distribution, and efficiency. The *background distribution* (blue, dotted) is constructed by analyzing a large number of (catalogs of) simulated GR sources and computing their log odds ratios. Given a choice of maximum tolerable false alarm probability β, this sets a *threshold* $\ln \mathcal{O}_\beta$ (the vertical dashed line) for the *measured* log odds ratio to overcome: $\ln \mathcal{O}_\beta$ is the value of log odds ratio such that a fraction β of the background is above it. For a given type of GR violation, one can also construct a *foreground distribution* of log odds ratio (red, dashed and solid). The *efficiency* ζ in finding the deviation from GR is the fraction of foreground above threshold (solid red).

For BBH (and possibly NSBH), one might consider using *e.g.* the very accurate inspiral-merger-ringdown waveform model with precessing spins that was recently presented by Pan *et al.* [32]. However, EOB-based waveforms tend to take orders of magnitude longer to generate than TaylorF2, making background calculations prohibitively expensive; on the other hand, this approximant might be extremely useful as an injection waveform for background calculations. For recovery, ideally one would want to have an analytic, frequency domain waveform which would still capture all the relevant physics: inspiral with precessing spins, merger, and ringdown. While the present paper was being finished, Hannam *et al.* [33] published what appears to be just such a waveform. This approximant, or a further improvement of it, may well be what is needed to extend TIGER to BBH. Alternatively, an upgrade of the fast time domain "PhenSpin" approximant of Sturani *et al.* could be used for recovery [34, 35]. However, for now we focus on BNS.

C. Selection of detections for analysis with TIGER

If, for now, we are going to restrict ourselves to BNS events for follow-up with TIGER, we need a way to distinguish between BNS detections on the one hand, and on the other hand NSBH and BBH events [51]. In recent years, the catalog of electromagneticallly observed binary pulsar systems has increased to the point where the neutron star mass distribution in binaries can be probed. For systems where both components are neutron stars, Valentim *et al.* [52] found that their masses peak at $\sim 1.37\,M_\odot \pm 0.042\,M_\odot$. Kiziltan *et al.* [53] restricted themselves to the 9 systems where the mass measurements are the most reliable, and found that the mass distribution peaks at $\mu_{\rm NS} \sim 1.34\,M_\odot$, having a width of $\sigma_{\rm NS} \sim 0.06\,M_\odot$. Now, in gravitational wave detection and parameter estimation, it is the *chirp mass* $\mathcal{M} = M\eta^{3/5}$ that is the most reliably determined, with uncertainties of a few percent[80]. Within the $2\,\sigma_{\rm NS}$ interval for m_1, m_2 found in [53], this quantity varies in the range $1.06 - 1.27\,M_\odot$. For NSBH and BBH systems, one must rely on theoretical models. In Dominik *et al.* [54], results from a large number of formation models for compact binaries are given. They find the minimum chirp mass for NSBH to be $1.7\,M_\odot$, and $2.4\,M_\odot$ for BBH. Thus, selecting only detections for which *e.g.* $\mathcal{M} < 1.3\,M_\odot$ at 95% confidence should remove all NSBH and BBH events. This would leave BNS systems with component masses up to $1.5\,M_\odot$, or $2.5\,\sigma_{\rm NS}$ above the mean of Kiziltan *et al.* [53]. Of course, it is entirely possible that some genuine BNS detections will be removed in this way (in fact, this is what the BNS results of [54] suggest), but the procedure is a conservative one.

Finally, we note that of necessity, the selection will have to be done based on parameter estimation with *GR waveforms*. If GR is incorrect, then there could be a large bias in the measurement of (among other parameters) $\mathcal{M}$ [22, 23, 27], in which case even a BBH system could be mis-classified as a BNS system. However, in that case we expect TIGER to *a fortiori* indicate a violation of GR.

D. The analysis pipeline

The TIGER analysis pipeline is part of LALInference, a software package within the LIGO Algorithm Library dedicated to Bayesian inference on gravitational wave detections [45]. In Secs. II A and II B, we explained the workings of the individual components; here we put everything together.

Assume that a number of compact binary coalescence detections have been made by the dedicated search pipelines [55–63]. Then the standard LALInference parameter estimation routines described in [45] will be used on each of them, with a variety of GR waveform models for recovery. On the basis of the measured values of chirp mass, BNS events can be selected to be followed up by TIGER; denote these by $\{d_1, d_2, \ldots, d_{\mathcal{N}}\}$. Subsequently, the TIGER pipeline will go through the following steps:

- For each detection $d_A \in \{d_1, d_2, \ldots, d_{\mathcal{N}}\}$, compute the Bayes factors ${}^{(A)}B^{i_1 i_2 \ldots i_k}_{\rm GR}$ for all the sub-hypotheses $H_{i_1 i_2 \ldots i_k}$ against $\mathcal{H}_{\rm GR}$ through Eqns. (6)–(10). Using Eq. (5), the *measured* combined log odds ratio for the catalog of detections, $\overline{\ln \mathcal{O}}^{\rm modGR}_{\rm GR}$, is obtained.

- For each detection d_A, perform a large number of GR injections with parameters picked from astrophysically motivated distributions (see Sec. III below), collectively denoted κ_A, in stretches of data surrounding d_A. This is to ensure that for each of the detections we have a set of simulated GR signals analyzed in a noise realization that will be as similar as possible to the one which the detection is in. For each of the injections, compute the ${}^{(A)}B^{i_1 i_2 \ldots i_k}_{\rm GR}$.

- Create a set $\kappa_{\rm cats}$ of simulated *catalogs* of $\mathcal{N}$ GR injections by randomly picking one simulated source from each of the injection sets κ_A, $A = 1, \ldots, \mathcal{N}$, making sure never to use the same injection twice. For each catalog, compute the combined log odds ratio $\ln \mathcal{O}^{\rm modGR}_{\rm GR}$. This will give us the *background distribution* $P(\ln \mathcal{O}|\mathcal{H}_{\rm GR}, \kappa_{\rm cats}, I)$.

- For a given maximum tolerable false alarm probability β, the background distribution can be used to set a threshold $\ln \mathcal{O}_\beta$, as explained in Sec. II B.

- Should it be the case that the measured log odds ratio is above threshold, $\overline{\ln \mathcal{O}}^{\rm modGR}_{\rm GR} > \ln \mathcal{O}_\beta$, then it will be of interest to compute the *actual* false alarm probability $\bar{\beta}$, given by

$$\bar{\beta} = \int_{\overline{\ln \mathcal{O}}^{\rm modGR}_{\rm GR}}^{\infty} P(\ln \mathcal{O}|\mathcal{H}_{\rm GR}, \kappa_{\rm cats}, I)\, d \ln \mathcal{O}. \qquad (13)$$

  This will then tell us to what extent the presence of a GR violation should be believed.

E. Performance of TIGER

The performance of the basic method summarized in Secs. II A and II B above was assessed in Li *et al.* [27, 28], where simulated signals from astrophysically distributed BNS sources with or without a GR violation were coherently added to stationary, Gaussian noise following the expected Advanced LIGO and Advanced Virgo final design sensitivities. The signals were randomly combined into catalogs of 15 sources each. These were analyzed as described above, with testing parameters $\{\psi_1, \psi_2, \psi_3\}$. An in-depth discussion of the results can be found in [27, 28]; in the interest of making the present discussion self-contained, we summarize the main conclusions.

Li *et al.* considered a variety of heuristic GR violations. It was found that a constant shift of 10% in the 1.5PN phase coefficient ψ_3 (which is of particular interest, since the dynamical non-linearities of gravity first appear at this order [6, 7]) could be seen with essentially 100% efficiency irrespective of maximum tolerable false alarm probability. Uncovering smaller shifts in ψ_3 (*e.g.* 2.5%) requires a larger number of sources per catalog. A shift of 20% in ψ_4 could also be found with high efficiency, despite that this parameter was not among the testing parameters. Deviations from the general phase structure in Eq. (1) were considered, such as the presence of an extra term at "1.25PN" order, *i.e.* $\propto f^{-5/6}$, and even a term with a mass dependent power of frequency, *i.e.* $\propto f^{\xi(M)}$ for some function $\xi(M)$, effectively ranging from 0.5PN to 1.5PN depending on total mass M. For the latter cases, the deviation could again be uncovered with almost 100% efficiency, on condition that the size of the violation was such that the phase at $f \sim 150$ Hz would differ from the GR phase by more than $\mathcal{O}(10)$ radians. Thus, we expect TIGER to be sensitive to a very wide range of possible GR violations.

As was already hinted at in [27–29], the method effectively circumvents potential problems related to a non-GR model being insufficiently parsimonious when the true number of additional parameters is in fact small. If a model has too many additional parameters compared to how many extra parameters are actually in the signal, then in *e.g.* Eq. (7), the likelihood function $p_{i_1 \ldots i_k}(d_A|\vec{\theta}, \delta\chi_{i_1}, \ldots, \delta\chi_{i_k}, I)$ might still be strongly peaked at the correct values of $\vec{\theta}$ and $\delta\chi_{i_1}, \ldots, \delta\chi_{i_k}$. However, the integration against the prior density $\pi_{i_1 \ldots i_k}(\delta\chi_{i_1}, \ldots, \delta\chi_{i_k}|I)$ over a parameter space whose dimensionality is too high might render the resulting evidence $P(d_A|H_{i_1 \ldots i_k}, I)$ small compared to $P(d_A|\mathcal{H}_{\rm GR}, I)$, even when GR is incorrect. In Bayesian studies preceding [27, 28], essentially only the hypothesis $H_{12 \ldots N_T}$ was compared with GR. One might expect that a GR violation will affect *all* the phase coefficients, in which case the latter would be the best thing to do, but this is not necessarily so; for instance, a non-zero graviton mass primarily modifies ψ_2. In *all* the examples considered by Li *et al.*, the most general sub-hypothesis $H_{12 \ldots N_T}$ was almost always disfavored compared to several sub-hypotheses will a smaller number of free parameters (see *e.g.* Figs. 6 and 15 of [27]). This includes the violation where a term of the form $f^{\xi(M)}$ was introduced. All this suggests that in the TIGER framework, the total number of testing parameters can be arbitrarily large without the results being impacted by insufficient parsimony in cases where the true number of non-GR parameters is small. However, this is contingent upon the background not being too much affected by the number of testing parameters: Waveform models with more free parameters will also more easily accommodate features in the noise, so that in principle the background could widen significantly as more testing parameters are allowed. In Sec. III E below, we show that in fact, the background is largely insensitive to an increase in the number of testing parameters used, in that it does not change significantly when going from *e.g.* three to four testing parameters.

Finally, TIGER is well-suited to the regime of low SNRs. This is because noise can make it difficult to discern the true nature of a GR violation, and even if the correct hypothesis is among the $H_{i_1 \ldots i_k}$, some other sub-hypothesis may still be favored. In [27], the example was given of a violation where H_3 was the correct hypothesis, but TIGER actually does better than a method which

only compares H_3 with $\mathcal{H}_{\rm GR}$, the reason being that occasionally some other sub-hypothesis will have a larger Bayes factor against GR. In this regard we note that the purpose of TIGER is not to identify the precise nature of a GR violation (in fact, the deviation from GR could be outside the family of violations considered in $\mathcal{H}_{\rm modGR}$), but first and foremost to establish whether or not one is present.

III. ROBUSTNESS AGAINST UNKNOWN FUNDAMENTAL, ASTROPHYSICAL, AND INSTRUMENTAL EFFECTS

Having outlined the structure of TIGER as a data analysis pipeline, we now show how it can be made robust against effects of a fundamental, astrophysical, or instrumental nature which can not easily be accounted for in our waveform models. In turn, we study the impact of neutron star tidal deformability, differences between waveform approximants, unknown contributions to the phase at high PN order, instrumental calibration errors, the effect on the background of the number of coefficients used, and precessing neutron star spins. We expressly gauge the importance of each these issues separately, in order to clearly demonstrate how each of them can be brought under control, before finally considering the situation where all of them are jointly present.

The results below pertain to simulations of BNS signals in stationary, Gaussian noise following the design sensitivity of Advanced LIGO and Virgo [8, 9]. Component masses were in the range $1-2\,M_\odot$,[81] sky positions and orientations were chosen from uniform distributions on the sphere, and sources were placed uniformly in comoving volume with luminosity distance $D \in [100, 250]$ Mpc. Depending on the type of robustness test, the signal waveform was taken to be TaylorF2 with zero or (anti-)aligned spins, or TaylorT4 with precessing spins; the recovery was done with TaylorF2 waveforms, again with either zero or (anti-)aligned spins. Only sources with optimal network SNR above 8 were taken into account [64]. Occasionally it would happen that a source survived the SNR cut without being found by the GR waveform model, meaning $\ln B^{\rm GR}_{\rm noise} \simeq 0$, with $\ln B^{\rm GR}_{\rm noise}$ the log Bayes factor for the hypothesis of a GR signal being present against the noise-only hypothesis. Such sources were discarded by imposing $\ln B^{\rm GR}_{\rm noise} > 32$, motivated by the fact that the main contribution to $\ln B^{\rm GR}_{\rm noise}$ is $(1/2)\,\langle h_{\rm GR}|h_{\rm GR}\rangle = (1/2)\,{\rm SNR}^2$, with $h_{\rm GR}$ the GR waveform, and $\langle\,\cdot\,|\,\cdot\,\rangle$ is the usual noise-weighted inner product [65]:

$$\langle a|b\rangle \equiv 4\Re\int_{f_0}^{f_{\rm LSO}} df\,\frac{\tilde a^*(f)\,\tilde b(f)}{S_n(f)}, \tag{14}$$

where a tilde denotes the Fourier transform, and $S_n(f)$ is the one-sided noise power spectral density. Finally, in the nested sampling process we used 1000 "live points" and 100 "MCMC points" (see [41–43] for definitions), which leads to an uncertainty $\lesssim 1$ in log Bayes factors against noise [43].

In what follows, we will want to compare different background distributions: with or without calibration errors, with or without tidal effects in the injections, and so on. A convenient way of quantifying the difference between distributions is by means of the Kolmogorov-Smirnov statistic [66, 67]. Consider backgrounds $P(\ln\mathcal{O}|\mathcal{H}_{\rm GR},\kappa_1,I)$ and $P(\ln\mathcal{O}|\mathcal{H}_{\rm GR},\kappa_2,I)$ for different injection sets κ_1, κ_2 (or in the case of calibration errors, different simulated data sets containing injections). Construct the *cumulative* distributions of log odds ratio and call these $F_{1,N}(\ln\mathcal{O})$ and $F_{2,N'}(\ln\mathcal{O})$, respectively; here N and N' are the numbers of log odds ratio values that are available in each of the two cases. Then the Kolmogorov-Smirnov (KS) statistic is just the largest distance between the cumulative distributions:

$$D^{1,2}_{N,N'} \equiv \sup_{\ln\mathcal{O}}|F_{1,N}(\ln\mathcal{O}) - F_{2,N'}(\ln\mathcal{O})|. \tag{15}$$

Note that by construction, this is a number between 0 and 1. If $D^{1,2}_{N,N'} \ll 1$, then the difference between the background distributions can be considered small.

A. Neutron star tidal deformability

As two neutron stars spiral towards each other, each will get deformed due to the tidal field of the other. These deformations have an influence on the orbital motion which gets imprinted onto the emitted gravitational wave signal. The size of the effect is set by the *tidal deformability* $\lambda({\rm EOS}, m)$, which relates the Newtonian tidal tensor $\mathcal{E}_{ij}$ of one star to the induced quadrupole moment Q_{ij} of the other: $Q_{ij} = -\lambda({\rm EOS}, m)\,\mathcal{E}_{ij}$. One has $\lambda(m) = (2/3)k_2(m)R^5(m)$, with k_2 the second Love number and R the neutron star radius. As the notation suggests, the tidal deformability depends on mass in a way that is determined by the neutron star equation of state (EOS). In the presence of tidal effects, the waveform phase takes the form $\Phi(v) = \Phi_{\rm PP}(v) + \Phi_{\rm tidal}(v)$, where $\Phi_{\rm PP}(v)$ is the usual point particle contribution, and to 1PN beyond leading order for tidal contributions one has[82] [68]

$$\Phi_{\rm tidal}(v) = \sum_{a=1}^{2} \frac{3\lambda_a}{128\eta M^5}\left[-\frac{24}{\chi_a}\left(1+\frac{11\eta}{\chi_a}\right)v^5 - \frac{5}{28\chi_a}\left(3179-919\,\chi_a-2286\,\chi_a^2+260\,\chi_a^3\right)v^7\right]. \quad (16)$$

The sum is over the components of the binary, and $\lambda_a = \lambda(m_a)$, $\chi_a = m_a/M$ for $a = 1, 2$. Note that although these contributions occur at 5PN and 6PN in the phase, they come with a prefactor that is potentially quite large: $\lambda/M^5 \propto (R/M)^5 \sim 10^2 - 10^5$ [69], so that the effect can be noticeable even with second-generation detectors. Indeed, in [70] it was shown that, if one assumes GR to be correct, the EOS can be significantly constrained by combining information from $\mathcal{O}(20)$ BNS observations. This in turn means that tidal effects could be mistaken for GR violations.

Since little is known about the EOS – in fact, currently the tidal deformability is uncertain by an order of magnitude – we have no way of including an accurate description of it in our waveform models. However, because of the high PN order at which these effects occur, they will only be important at very high frequencies. Indeed, as shown by Hinderer *et al.* [68] (see also the recent work by Read *et al.* [71]), with second-generation detectors they only become noticeable for $f > 450$ Hz. For this reason we terminate our template waveforms at $f = 400$ Hz (which in terms of characteristic velocity and compactness corresponds to $v/c \sim 0.25$ and $GM/(c^2R) \sim 0.07$, respectively). As it turns out, this leads to a loss in SNR of less than a percent, and in any case TIGER mostly probes the lower PN orders, corresponding to lower frequencies. However, here too we want to explicitly check that this suffices to make TIGER impervious to the unknown effect.

In Fig. 2, we compare the background for TaylorF2 injections without tidal effects, with the background obtained from injections with a very hard EOS (corresponding to large deformability), namely the one labeled MS1 in [68]. The injected waveforms are taken to terminate at LSO while the recovery waveforms (also TaylorF2) are cut off at 400 Hz in both cases.

Consider the background distribution for "point particle" (PP) injections $\kappa_{\rm PP}$ (no tidal effects), $P(\ln\mathcal{O}|\mathcal{H}_{\rm GR}, \kappa_{\rm PP}, I)$, and the distribution of log odds ratio for MS1 injections $\kappa_{\rm MS1}$, $P(\ln\mathcal{O}|\mathcal{H}_{\rm GR}, \kappa_{\rm MS1}, I)$. Using the cumulative distributions of log odds in the two cases, one can construct the Kolmogorov-Smirnov statistic as in Eq. (15). For the injection sets $\kappa_{\rm PP}$, $\kappa_{\rm MS1}$ used in Fig. 2, we find $D^{\rm PP,MS1}_{N,N'} = 0.06$, indicating that the two background distributions are very close to each other. We conclude that the 400 Hz cut-off renders tidal effects invisible without affecting TIGER's ability to look for GR violations.

Below we will continue to implement a 400 Hz cut-off in the recovery waveforms.

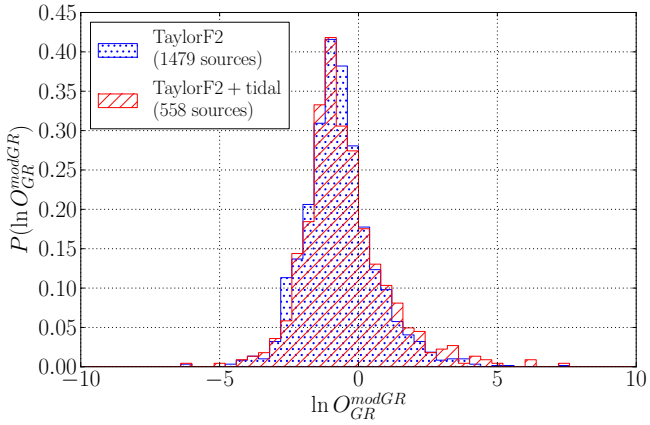

FIG. 2: Single-source background distributions for TaylorF2 injections without tidal effects (blue, dotted) and injections with strong tidal deformability (red, dashed), both analyzed with TaylorF2 waveforms that are cut off at $f = 400$ Hz.

B. Differences between waveform approximants

For all the post-Newtonian waveform approximants, the phase $\phi(t)$ and instantaneous velocity $v(t)$ (or equivalently $t(v)$) are computed from the conserved energy per unit mass $E(v)$ and the gravitational wave flux $\mathcal{F}(v)$ through Kepler's law and the flux-energy balance equation:

$$\frac{d\phi}{dt} - \frac{v^3}{M} = 0, \quad (17)$$

$$\frac{dv}{dt} + \frac{\mathcal{F}}{ME'(v)} = 0, \quad (18)$$

where the prime denotes derivation with respect to v. The solutions take the general form

$$t(v) = t_{\rm ref} + M\int_v^{v_{\rm ref}} dv\,\frac{E'(v)}{\mathcal{F}(v)}, \quad (19)$$

$$\phi(v) = \phi_{\rm ref} + \int_v^{v_{\rm ref}} dv\,v^3\frac{E'(v)}{\mathcal{F}(v)}, \quad (20)$$

where $t_{\rm ref}$ and $\phi_{\rm ref}$ are integration constants, and $v_{\rm ref}$ is an arbitrary reference velocity. Now, since $E(v)$ and $\mathcal{F}(v)$ are known as series expansions in v up to a finite order, there are multiple ways of treating the above equations. In the case of the so-called TaylorT1 approximant, $E'(v)/\mathcal{F}(v)$ is kept as a ratio of polynomials, and Eqns. (17) and (18) are solved numerically. In the case of TaylorT4, what one does instead is to expand the ratio $E'(v)/\mathcal{F}(v)$ and truncate the result at the consistent PN

order, after which Eqns. (17), (18) are again solved numerically. TaylorT2 is obtained by expanding and consistently truncating $E'(v)/\mathcal{F}(v)$, and integrating Eqns. (19) and (20) to obtain a pair of transcendental equations for ϕ and t as functions of v, which are then solved numerically. For TaylorT3 one also expands and truncates $E'(v)/\mathcal{F}(v)$, and integrates Eqns. (19) and (20) to obtain expressions for $\phi(v)$ and $t(v)$. The latter is inverted to $v(t)$, and a representation of $\phi(t) = \phi(v(t))$ is computed. Finally, the frequency domain TaylorF2 approximant is obtained through the *stationary phase approximation*, by utilizing a saddle point in the calculation of the Fourier transform of the time domain waveform. For more details on all these approximants, see [31] and references therein.

A qualitatively different way of obtaining waveform models is the effective-one-body (EOB) method. Here a mapping is established between the motion of the two component masses and the motion of a *single* particle in an effective metric, which is captured by a set of Hamiltonian equations for the angular and radial motion. These are solved numerically. The advantage of this method is that the resulting waveforms are reliable up to later times compared to the PN ones (well into the plunge preceding merger), which also means that they lend themselves particularly well to being further "tuned" using input from numerical simulations after being completed with a ringdown waveform. Here too we point to [31] and references therein for further information.

The authors of [31] calculated the *effectualness* and *faithfulness* of post-Newtonian waveforms with respect to each other, as well as with an EOB waveform model tuned using numerical simulations, and this for a variety of component masses. The effectualness is a measure of how effective a waveform model h_t will be when used as a template to detect a "signal" waveform h_s; for given intrinsic and extrinsic signal parameters $\vec{\lambda}$, it is defined as $\max_{\vec{\theta}} \langle \hat{h}_s(\vec{\lambda}) | \hat{h}_t(\vec{\theta}) \rangle$, where $\hat{h} \equiv h/\sqrt{\langle h|h \rangle}$, and $\langle \cdot | \cdot \rangle$ is again the usual noise-weighted inner product [65]. In the case of faithfulness, the intrinsic parameters $\vec{\lambda}_{\rm intr}$ of "signal" and "template" are taken to be the same, and the maximization is only over the template's time and phase at coalescence: $\max_{t_c,\varphi_c} \langle \hat{h}_s(\vec{\lambda}_{\rm intr}) | \hat{h}_t(\vec{\lambda}_{\rm intr}) \rangle$. In the expected mass range of NSBH and BBH, there can be significant differences between the PN approximants amongst themselves, and with EOB waveforms. However, in the BNS mass range, at least in the case of zero spins, both the effectualness and faithfulness for any pair of PN waveforms and for any PN approximant with the EOB model tend to be above 0.99.[83] For example, in the case of Advanced LIGO and for $(m_1, m_2) = (1.42, 1.38)\, M_\odot$, the faithfulness of TaylorF2 against TaylorT4 is 0.999, and for TaylorF2 against EOB it is 0.996.

The strong agreement between the various waveform approximants in the BNS mass range suggests that, at least for such systems, it is safe to adopt TaylorF2, the computationally least expensive waveform model, for the trial waveforms used in TIGER. However, since the pipeline is specifically meant to find small anomalies in the signals, we need to make sure that even small differences between waveform approximants are not mistaken for violations of GR.

In Fig. 3, we compare single-source background distributions for the case where the GR signals are TaylorT4 waveforms and the case where they are TaylorF2 waveforms; but, in both cases, the analysis of the data is done with TaylorF2. Once again the difference between the two distributions can be quantified by using the KS statistic, which in this case comes out to be $D^{\rm TF2,TT4}_{N,N'} = 0.07$.

Due to computational cost, we decided not to repeat the calculation with TaylorT1, TaylorT2, TaylorT3, or EOB injections. However, the results of Fig. 3, together with the waveform comparisons of [31], are sufficient to conclude that TIGER will not mistake differences in waveform models for violations of GR.

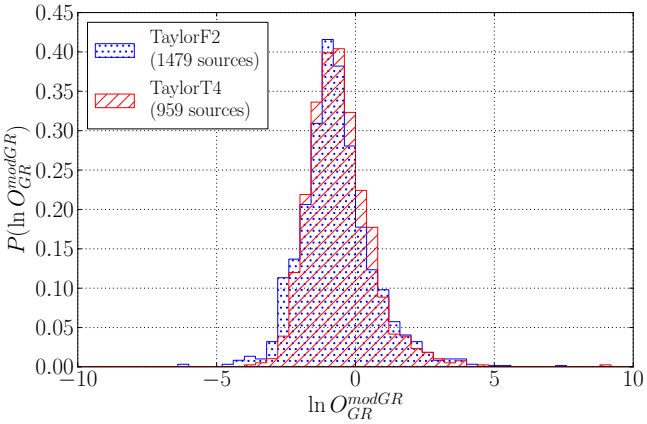

FIG. 3: Single-source background distributions for TaylorF2 injections (blue, dotted) and TaylorT4 injections (red, dashed), both analyzed with TaylorF2 waveforms cut off at 400 Hz.

C. Effect of post-Newtonian order

In [31], waveform approximants were considered up to 3.5PN in phase, which is the highest post-Newtonian order currently available. To this order, post-Newtonian waveforms and EOB-based inspiral-merger-ringdown waveforms that were tuned using numerical relativity simulations agree extremely well in the BNS mass regime. However, taking numerical relativity results to be the benchmark for how realistic a waveform model is, we note that large-scale numerical simulations of spacetimes containing coalescing binaries still only give information about the last few tens of cycles [72], whereas a typical BNS waveform is thousands of cycles long. Thus, it could be that adequate modeling of the signals by post-Newtonian waveforms will require going to still higher PN

order in the phase.[84]

In Fig. 4, we probe the effect on the background of differences in post-Newtonian order between signal and recovery waveforms, for TaylorF2. In one case, both are taken to 3.5PN order, while in the other case the signal is 3.5PN whereas the recovery waveform only goes to 3PN. We see that the distributions barely differ; the KS statistic is $D^{\rm 3PN,3.5PN}_{N,N'} = 0.05$. Needless to say, this does not prove that missing post-Newtonian orders beyond 3.5PN will be unproblematic, but it does lend further confidence to the soundness of our approach. Note also that one can expect high PN contributions to manifest themselves at high frequencies, and our recovery waveforms are cut of at 400 Hz.

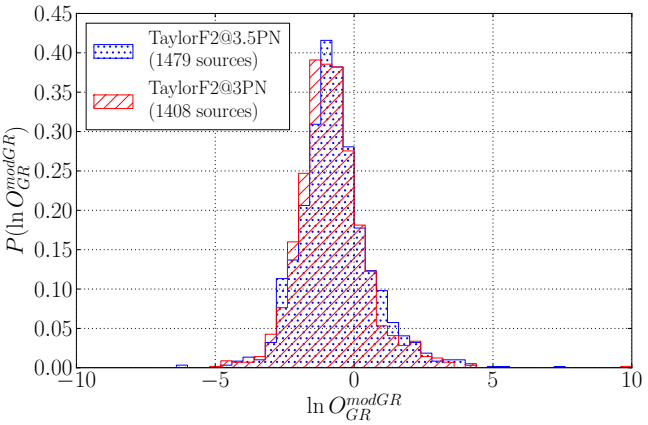

FIG. 4: Single-source background distributions for TaylorF2 injections to 3.5PN, where in one case the recovery waveform is also TaylorF2 to 3.5PN (blue, dotted) and in the other case, TaylorF2 to 3PN (red, dashed), both cut off at 400 Hz.

D. Instrumental calibration errors

Imperfect calibration of the instruments can cause one to draw incorrect conclusions about detected signals. Calibration errors affect the instruments' transfer functions $R(f)$, which relate external length changes $\Delta L_{\rm ext}$ in the interferometer arms to the detector outputs $e(f)$:

$$\Delta L_{\rm ext}(f) = R(f)\, e(f). \tag{21}$$

$R(f)$ is a complex function, which can be written in polar form as

$$R(f) = \left[1 + \frac{\delta A}{A}(f)\right] e^{i\delta\phi(f)}\, R_e(f), \tag{22}$$

where $(\delta A/A)(f)$ and $\delta\phi(f)$ are frequency dependent calibration errors in amplitude and phase, respectively, and $R_e(f)$ is the transfer function in the absence of errors. The frequency domain data stream is given by

$$\tilde{d}(f) = \frac{\Delta L_{\rm ext}(f)}{L}, \tag{23}$$

where L is the interferometer arm length in the absence of disturbances. In the expressions for the likelihood functions, Eq. (9), calibration errors enter both the data stream $\tilde{d}$ and the power spectral density of the noise $S_n(f)$, but *not* the model waveforms $\tilde{h}_{i_1\ldots i_k}$ and $\tilde{h}_{\rm GR}$, which is how parameter estimation and model selection get affected by them.

In [74], the calibration errors were modeled based on the errors measured in the initial LIGO and Virgo instruments, and their effect on Bayesian parameter estimation and model selection for advanced detectors was assessed. It was found that even with amplitude errors of $\delta A/A \sim 10\%$ and phase errors $\delta\phi \sim 3$ degrees in each instrument, for 90% of sources the systematics induced will be less than 20% of the statistical uncertainties in parameter estimation. Similarly, model selection is not much affected by calibration errors.

Fig. 5 shows the effect of calibration errors, modeled exactly as in [74], on the log odds ratio background distribution. As expected, the effect is minor (with $D^{\rm cal,nocal}_{N,N'} = 0.04$), and calibration errors will not affect the performance of TIGER.

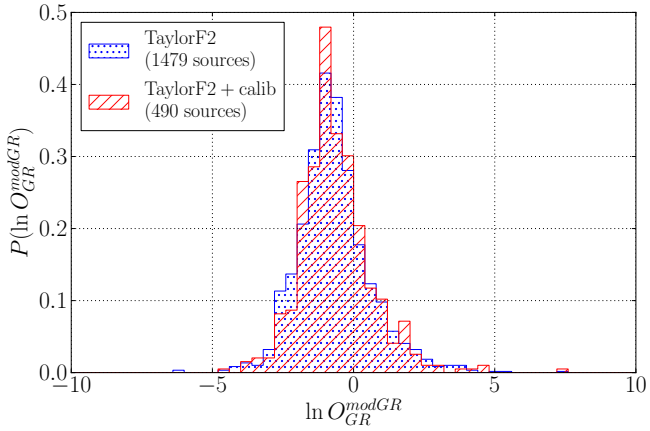

FIG. 5: Single-source background distributions for TaylorF2 injections without calibration errors (blue, dotted) and with frequency dependent amplitude and phase errors modeled as in [74] (red, dashed), both analyzed with TaylorF2 waveforms cut off at 400 Hz.

E. Number of testing parameters

As explained above, TIGER allows one to circumvent the usual problem in Bayesian analysis when the number of extra parameters in the model is too large: The total number of testing parameters, N_T, can in principle be arbitrarily large without risk of being penalized by the high dimensionality of the parameter space should the number of extra parameters in the signal be smaller than N_T. One aspect of this was already illustrated in [27], where it was shown that if the GR violation is limited to *e.g.* the 1.5PN phase coefficient, hypotheses with

too many free parameters tend to be disfavored even if they include ψ_3. However, what also needs to be checked explicitly is how sensitive the background is to the number of testing parameters: Should it be the case that it widens dramatically as N_T is increased because features in the noise can more easily be accommodated by waveforms with more free parameters, then the advantage disappears. In Fig. 6, we compare backgrounds for $N_T = 3$ and $N_T = 4$, and the difference turns out to be small; in terms of a KS statistic, $D^{3,4}_{N,N'} = 0.11$.

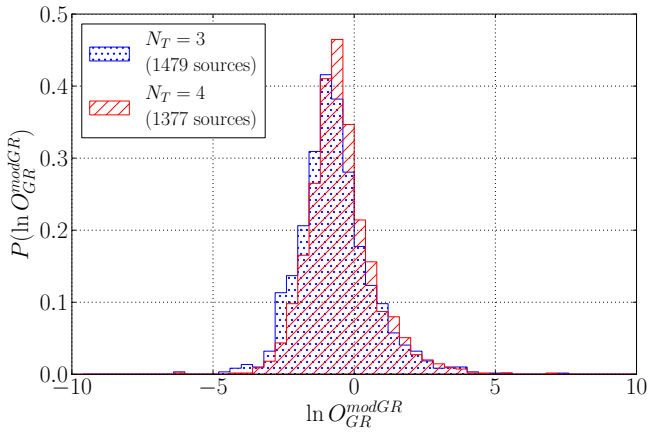

FIG. 6: Single-source background distributions for TaylorF2 injections with TaylorF2 recovery, in one case with three testing parameters (blue, dotted) and in the other with four (red, dashed).

Together with the results of [27], this indicates that one should use as many testing parameters as possible. However, in practice there will be computational constraints due to the exponential growth of the number of sub-hypotheses with the total number of testing parameters; indeed, for N_T testing parameters, $2^{N_T} - 1$ sub-hypotheses $H_{i_1 \ldots i_k}$ need to be compared with $\mathcal{H}_{\rm GR}$. The results of [27, 28] suggest that in the case of BNS, the sensitivity of TIGER to GR violations occurring above 2PN order in phase will be limited. In the examples below, we use three testing parameters, $\{\psi_1, \psi_2, \psi_3\}$.

F. Neutron star spins

The observed pulsar spin periods and assumptions about neutron star spindown rates lead to periods at birth in the range 10–140 ms [46], corresponding to dimensionless spins $J/m^2 \lesssim 0.04$, and the fastest known pulsar in a BNS has a spin $J/m^2 \sim 0.02$. Thus, neutron star spins in BNS systems are generally expected to be small. Nevertheless, we need to quantify their effect on the background distribution and hence the detectability of GR violations.

In the phase, spin-orbit effects first appear at 1.5PN order, and spin-spin effects at 2PN. The amplitude is also affected, primarily because of spin-induced precession of the orbital plane, which causes the inclination angle to change so that sometimes a system might be close to being face-on whereas at other times it will be closer to being edge-on, causing amplitude modulation.

To describe the orbital motion with inclusion of spins, one again uses the Kepler and flux-energy balance equations, Eq. (17)–(18), with $E(v)$ and $\mathcal{F}(v)$ modified to take spin-orbit and spin-spin effects into account, and these are supplemented by a set of differential equations for the time evolution of the individual spins $\vec{S}_1$ and $\vec{S}_2$, and of the unit normal in the direction of orbital angular momentum, $\hat{L}$. For the purposes of this paper, spin effects were included to 2.5PN [47], although by now spin-orbit effects in the flux are known to 3.5PN [48]. In the case of spins that are (anti-)aligned with each other and the orbital angular momentum, so that there is no precession, it is not difficult to arrive at a closed expression for phase as a function of frequency in the stationary phase approximation [49].

To assess the effect of spins, we constructed a background where the injected signals were TaylorT4 waveforms with precessing spins included in the dynamics, as described above. (Results for injections with (anti-)aligned spinning TaylorF2 waveforms were already reported in [50].) The spin orientations were picked from a uniform distribution on the sphere, and their magnitudes followed a Gaussian distribution centered on zero and with $\sigma = 0.05$. The recovery waveforms were again TaylorF2, but this time allowing for spins that are aligned or anti-aligned with orbital angular momentum. We need to pick a prior distribution for the spin magnitudes in the recovery waveform. In the present setting, the most natural choice is again a Gaussian centered on zero and having a width of 0.05. Indeed, letting spins in the recovery waveform vary within a wide range could lead us to miss GR violations occuring from 1.5PN order onward, since such deviations could be accomodated by adjusting the spins.

We explicitly note that the smallness of neutron star spins is an astrophysical assumption that enters the background calculation; see Sec. IV below for a discussion. However, given general astrophysical considerations as well as currently observed binary neutron star systems [46], most likely our choice of spin distributions in injections and recovery waveforms leads to a background that is rather conservative.

Since the injections have precessing spins while in the recovery we only allow for (anti-)aligned spins, the recovery waveform model will not perfectly capture the signal even for BNS. Nevertheless, the effect on the background distribution is minor, as shown in Fig. 7; one has $D^{\rm align,prec}_{N,N'} = 0.08$. Clearly, given the relative smallness of the spins, allowing for (anti-)aligned spins in TaylorF2 is sufficient for this waveform model to capture the spin effects in the signal, at least to the extent that the background is not significantly affected.

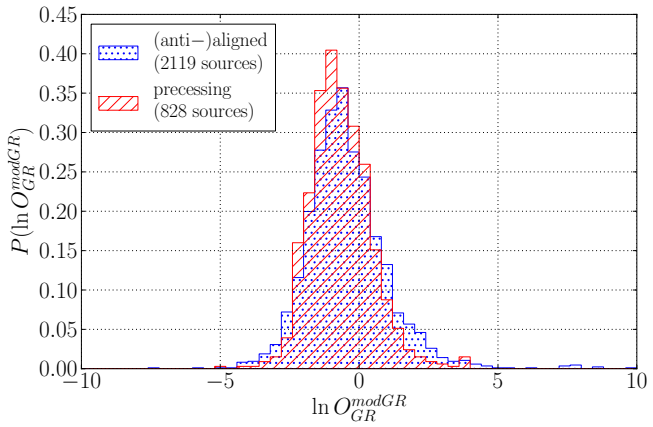

FIG. 7: Single-source background distributions for TaylorF2 injections with (anti-)aligned spins (blue, dotted) and TaylorT4 injections with precessing spins (red, dashed). In both cases the recovery is with TaylorF2 waveforms cut off at 400 Hz.

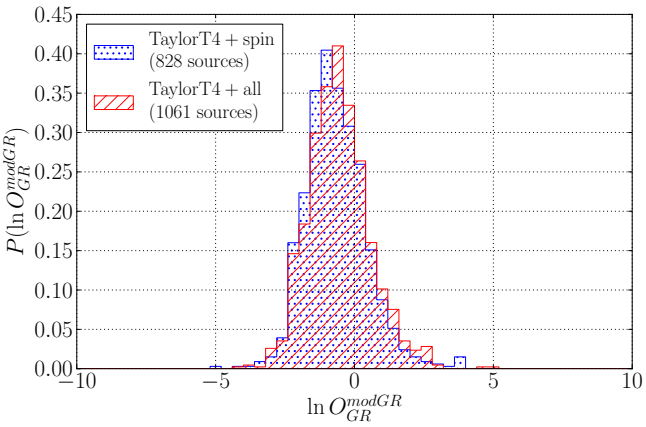

FIG. 8: Single-source background distributions for TaylorT4 injections with precessing spins (blue, dotted) and TaylorT4 injections with precessing spins, tidal effects, and calibration errors (red, dashed). In both cases, the recovery is with (anti-)aligned spinning TaylorF2 cut off at 400 Hz.

G. Combined effect of differences between waveform approximants, tidal deformation, calibration errors, and spins

We now put everything together and compute a background distribution where the recovery waveform is TaylorF2 with (anti-)aligned spins, cut off at 400 Hz, but the injections are TaylorT4 with precessing spins and tidal effects at 0PN and 1PN, and calibration errors are also included. In the case of TaylorT4, the phase is only computed numerically, and tidal effects must be added in the equation for $dv/dt(v)$:

$$\frac{dv}{dt}(v) = \mathcal{G}_{\rm PP}(v) + \mathcal{G}_{\rm tidal}(v), \tag{24}$$

where to 1PN order [76]

$$\begin{aligned}&\mathcal{G}_{\rm tidal}(v)\\ &= \frac{16\chi_1\lambda_2}{5M^6}\left[12(1+11\chi_1)\,v^{19}\right.\\ &\left.+\left(\frac{4421}{28}-\frac{12263}{28}\,\chi_2+\frac{1893}{2}\,\chi_2^2-661\,\chi_2^3\right)\,v^{21}\right]\\ &+(1\leftrightarrow 2).\end{aligned} \tag{25}$$

For the expression of the point particle contribution $\mathcal{G}_{\rm PP}(v)$ to 3.5PN, with spins included up to 2.5PN, we refer to [47].

For the case of single sources, the effect on the background of a combination of precessing spins, tidal effects, and calibration errors is shown in Fig. 8. In terms of a KS statistic, the difference between backgrounds is $D_{N,N'}^{\rm spins,all} = 0.07$.

For reasons of computational expense, so far we have only shown differences between backgrounds for *single sources*, which is appropriate for the case where there is only one detection. If there are $\mathcal{N}$ detections that can be clearly identified as BNS events according to the criterion of Sec. II C, then one will want to construct a background distribution for *catalogs* of $\mathcal{N}$ sources each, as explained in Sec. II D. We computed backgrounds using the injection sets of Fig. 8, but now randomly combining injections into catalogs of 15 sources each. The results are shown in Fig. 9. When information from multiple GR sources is combined, one expects $\mathcal{H}_{\rm GR}$ to be much more favored over $\mathcal{H}_{\rm modGR}$, and this is what we see: in both cases, the distribution of $\ln \mathcal{O}_{\rm GR}^{\rm modGR}$ stretches to much more negative values. However, when making comparisons of different physical set- ups, combining information from multiple sources can make the differences show up much more clearly than in the case of single sources. For the purposes of this paper, a much smaller number of simulations were performed than one would in reality; one has ${}^{\rm (cat)}D_{N,N'}^{\rm spins,all} = 0.24$, but this will in large part be due to small number statistics. Reassuringly, even for catalogs of sources, the two background distributions are rather similar, with both favoring strongly negative values of log odds.

Finally, we want to show at least one example of how well violations of GR might be detectable in the presence of strong tidal effects, instrumental calibration errors, and precessing spins. Recalling that the 1.5PN contribution to the orbital motion is where, according to GR, the dynamical self-interaction of spacetime first becomes visible [6, 7], we consider a (heuristic) violation of GR at that order, taking the form of a -10% shift in the relevant coefficient in the expansion of $dv/dt(v)$:

$$\begin{aligned}\frac{dv}{dt}(v) \;=\; &\mathcal{G}_{\rm PP}(v) + \mathcal{G}_{\rm tidal}(v)\\ &+\delta\xi_3\,\alpha_3(m_1,m_2,\vec{S}_1,\vec{S}_2)\,v^{12},\end{aligned} \tag{26}$$

where we note that the leading-order contribution to dv/dt goes like v^9; $\alpha_3(m_1,m_2,\vec{S}_1,\vec{S}_2)$ is the 1.5PN co-

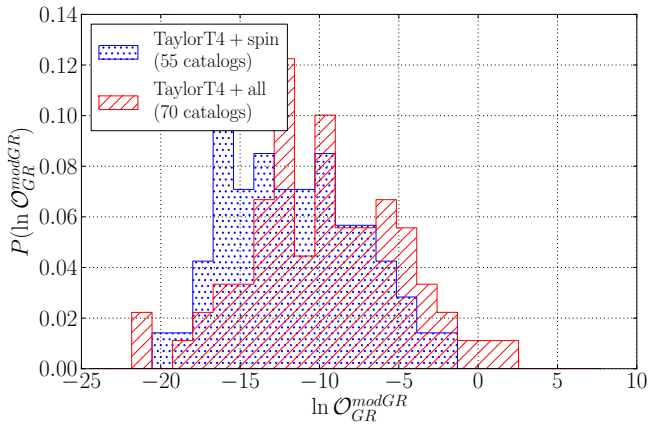

FIG. 9: The same comparison as in Fig. 8, but now for *catalogs* of 15 sources each. Note how GR is typically much more favored when information from multiple GR sources is combined.

efficient predicted by GR, and $\delta\xi_3 = -0.1$.

In Fig. 10, we show background as well as foreground log odds ratio distributions, for catalogs of 15 sources each, where in both cases the injections include neutron star tidal deformation, instrumental calibration errors, and precessing spins. As before, the recovery is with TaylorF2 waveforms that allow for (anti-)aligned spins, cut off at a frequency of 400 Hz. We see that the separation between the distributions is complete: almost regardless of false alarm probability, with 15 BNS detections the efficiency in finding the given GR violation is essentially 100%.

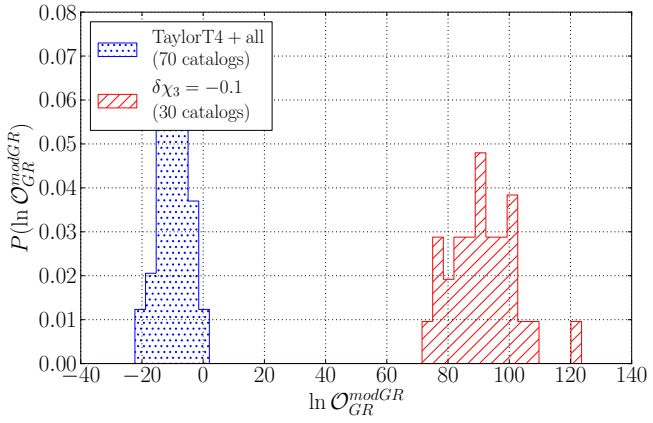

FIG. 10: Log odds ratio distributions for catalogs of 15 sources each. The blue, dotted histogram is the GR background for TaylorT4 signals with precessing spins, neutron star tidal deformation, and instrumental calibration errors. The red, dashed one is a foreground distribution for signals with the same effects present, and with a GR violation that takes the form of a constant -10% shift at 1.5PN, as explained in the main text. In both cases, the recovery is with (anti-)aligned spinning TaylorF2 waveforms cut off at 400 Hz.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

We have developed TIGER, a data analysis pipeline to perform model-independent tests of general relativity in the strong-field regime, using detections of compact binary coalescence events with second-generation gravitational wave detectors. In present form, it can already be applied to binary neutron star events, where waveform models that are reliable and can be generated sufficiently fast on a computer are available. The basic idea is to compare the GR hypothesis $\mathcal{H}_{\rm GR}$ with the hypothesis $\mathcal{H}_{\rm modGR}$ that one or more coefficients in the post-Newtonian expression for the phase do not depend on component masses and spins in the way GR predicts. Though the latter hypothesis has no waveform model associated with it, it can be written as the logical union of mutually exclusive sub-hypotheses, in each of which a fixed number of phase coefficients are free parameters on top of component masses, spins, sky position, orientation, and distance, while the others depend on masses and spins in the way GR predicts.

After selecting for BNS events on the basis of chirp mass, which can be reliably measured, TIGER computes Bayes factors against GR for all the sub-hypotheses. This is done for each selected event separately, after which all the information gathered is combined into a single quantity, the log odds ratio $\ln \mathcal{O}_{\rm GR}^{\rm modGR}$ for $\mathcal{H}_{\rm modGR}$ against $\mathcal{H}_{\rm GR}$. Next, a background distribution for the log odds ratio is constructed by injecting large numbers of simulated GR signals in stretches of data near the ones where the detections were made, which are assembled into catalogs, for each of which the log odds ratio is computed. With a choice of maximum tolerable false alarm probability, the background can be used to set a threshold for the *measured* log odds ratio to overcome.

TIGER is able to uncover a wide variety of GR violations, including ones which can in principle not be accommodated by $\mathcal{H}_{\rm GR}$. The total number of "testing parameters" N_T can be made arbitrarily large, subject only to computational constraints, as the availability of sub-hypotheses with any number of free parameters up to N_T allows one to circumvent the usual problems in Bayesian analysis in cases where the true number of additional parameters needed is in fact small. Noise can lead us to incorrectly identify the nature of a GR violation, so that the use of many sub-hypotheses increases our chances of finding a violation if one is present; this makes TIGER well-suited to the regime of low SNR detections. Finally, the ability to combine information from multiple sources leads to a stronger test.

We performed a range of numerical experiments to check the robustness of TIGER against fundamental, astrophysical, and instrumental unknowns. In the BNS mass regime, the differences between the available waveform approximants are very small, making it unlikely that imperfect modeling of the signal will cause us to suspect a violation of GR. The fact that waveforms are only

known up to a finite post-Newtonian order should also not be cause for concern. In the final stages of inspiral, finite size effects are important and the neutron stars will deform each other in an essentially unknown way; however, if the recovery waveforms are cut off at 400 Hz then the unknown tidal effects will not be mistaken for violations of GR, but the performance of TIGER remains unaffected. Instrumental calibration errors of expected size will not be problematic. Finally, if, as generally expected, the spins of neutron stars in binaries are small, then they can easily be dealt with. The next step will be to study the the behavior of TIGER in real noise, which is not quite stationary or Gaussian. We will test the pipeline using existing data taken by the initial LIGO and Virgo detectors, but "recolored" so that the underlying power spectral densities are the ones predicted for the advanced interferometers, while retaining the non-stationarities in the noise. Results will be reported in a forthcoming publication.

In present form, TIGER relies on two important astrophysical assumptions. One is that NSBH and BBH coalescences have chirp masses above a certain value, so that such events can be discarded, leaving only BNS. The other is the relative smallness of spins for BNS. In the future we will also want to work with BBH and NSBH events so that if an anomaly is discovered in BNS signals, we can confirm that it is of a fundamental rather than an astrophysical nature by using qualitatively different systems. Pan *et al.* appear to have arrived at a reliable semi-analytic waveform model for BBH and NSBH coalescence [32], and their approximant will be extremely useful as an injection waveform. However, it is too computationally expensive to be used for recovery. On the other hand, very recently Hannam *et al.* [33] proposed a *frequency domain* inspiral-merger-ringdown waveform which captures precessing spins, and which may already be useful for our purposes. An upgrade of the fast time domain "PhenSpin" waveform of Sturani *et al.* could also be an option for recovery [34, 35]. (Note that for the background calculation, it is important that the *injected* waveform model be as close as possible to reality, but the requirements for the recovery waveform are less stringent.) To have some idea of what might conceivably be possible with BBH, we used the earlier BBH waveform approximant of [79] with spins set to zero, for both injection and recovery, choosing component masses to be in the range $[5, 15]\,M_\odot$ and placing sources uniformly in co-moving volume with distances up to 1.25 Gpc. It was found that for catalogs of 20 sources each, a deviation in (the equivalent of) the 3PN phase coefficient ψ_6 of only 0.5% could be picked up with essentially 100% efficiency, using only $\{\psi_1, \psi_2, \psi_3, \psi_4\}$ as testing coefficients; see Fig. 6 of [29]. Here some caution is called for, considering that astrophysical black holes are likely to have large, non-aligned spins, but the result is encouraging. The possibility of reliably applying TIGER to BBH detections using a waveform model along the lines of Hannam *et al.* [33] or Sturani *et al.* [34, 35] will be a subject of intense investigation.

With the construction of TIGER, the problem of *finding* a deviation from GR with second-generation detectors is essentially solved, at least for the case of BNS. A still open problem is to identify the *nature* of a GR violation should one be present. Here one could compute posterior densities for a number of free parameters, not necessarily restricted to post-Newtonian phase parameters but possibly using the more general PPE waveforms [23, 25]. However, whatever model is used, it remains the case that the inferred values of the non-GR parameters can be affected by fundamental bias [22, 23] if the true signal is not included in the signal model, even in the high SNR limit (see the examples in [27]). The problem of determining whether the modeled deviations adequately explain the observations without bias is left for future studies.

Acknowledgements

MA, WDP, TGFL, CVDB, and JV were supported by the research programme of the Foundation for Fundamental Research on Matter (FOM), which is partially supported by the Netherlands Organisation for Scientific Research (NWO). SV acknowledges the support of the National Science Foundation and the LIGO Laboratory. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation and operates under cooperative agreement PHY-0757058. The authors would like to acknowledge the LIGO Data Grid clusters, without which the simulations could not have been performed. Specifically, these include the computing resources supported by National Science Foundation awards PHY-0923409 and PHY-0600953 to UW-Milwaukee. Also, we thank the Albert Einstein Institute in Hannover, supported by the Max-Planck-Gesellschaft, for use of the Atlas high-performance computing cluster. It is a pleasure to thank E. Berti, A. ter Braack, A. Buonanno, N. Cornish, J.D.E. Creighton, W.M. Farr, B.R. Iyer, C.K. Mishra, C. Pollice, B.S. Sathyaprakash, R. Sturani, and N. Yunes for useful discussions.

[1] C.W. Misner, K.S. Thorne, and J.A. Wheeler, *Gravitation*, W.H. Freeman and Company, New York, 1973

[2] C. M. Will, Liv. Rev. Rel. **9**, 3 (2006) `http://www.livingreviews.org/lrr-2006-3`

[3] R.A. Hulse and J.H. Taylor, Astrophys. J. **195**, L51 (1975)

[4] P.C.C. Freire *et al.*, Mon. Not. R. Astron. Soc. **423**, 3328 (2012)

[5] Antoniadis *et al.*, Science **340**, 6131 (2013)
[6] L. Blanchet and B.S. Sathyaprakash, Class. Quantum Grav. **11**, 2807 (1994)
[7] L. Blanchet and B.S. Sathyaprakash, Phys. Rev. Lett. **74**, 1067 (1995)
[8] G.M. Harry for the LIGO Scientific Collaboration, Class. Quantum Grav. **27**, 084006 (2010)
[9] Advanced Virgo Baseline Design, The Virgo Collaboration, technical note VIR027A09 (2009)
[10] H. Grote (for the LIGO Scientific Collaboration), Class. Quantum Grav. **25**, 114043 (2008); H Grote and the LIGO Scientific Collaboration, Class Quantum Grav. **27**, 084003 (2010)
[11] K. Kuroda for the LCGT Collaboration, Class. Quantum Grav. **27**, 084004 (2010)
[12] C.S. Unnikrishnan, to appear in Int. J. Mod. Phys D; LIGO Scientific Collaboration technical note P1200166-v1 (2012)
[13] LIGO Scientific Collaboration, Virgo Collaboration, Class. Quantum Grav. **27**, 173001 (2010)
[14] LIGO Scientific Collaboration, Virgo Collaboration; arXiv:1304.0670 [gr-qc]
[15] H.-Y. Chen and D.E. Holz; arXiv:1206.0703 [astro-ph]
[16] J.R. Gair, M. Vallisneri, S.L. Larson, J.G. Baker, to appear in Liv. Rev. Rel.; arXiv:1212.5575 [gr-qc]
[17] N. Yunes and X. Siemens, to appear in Liv. Rev. Rel.; arXiv:1304.3473 [gr-qc]
[18] K.G. Arun, B.R. Iyer, M.S.S. Qusailah, and B.S. Sathyaprakash, Class. Quantum Grav. **23**, L37 (2006)
[19] K.G. Arun, B.R. Iyer, M.S.S. Qusailah, and B.S. Sathyaprakash, Phys. Rev. D **74**, 024006 (2006)
[20] C.K. Mishra, K.G. Arun, B.R. Iyer, and B.S. Sathyaprakash, Phys. Rev. D **82**, 064010 (2010)
[21] S. Gossan, J. Veitch, and B.S. Sathyaprakash, Phys. Rev. D **85**, 124056 (2012)
[22] W. Del Pozzo, J. Veitch, and A. Vecchio, Phys. Rev. D **83**, 082002 (2011)
[23] N. Yunes and F. Pretorius, Phys. Rev. D **80**, 122003 (2009)
[24] N. Cornish, L. Sampson, N. Yunes, and F. Pretorius, Phys. Rev. D **84**, 062003 (2011)
[25] K. Chatziioannou, N. Yunes, and N. Cornish, Phys. Rev. D **86**, 022004 (2012)
[26] L. Sampson, N. Yunes, and N. Cornish; arXiv:1307.8144 [gr-qc]
[27] T.G.F. Li *et al.*, Phys. Rev. D **85**, 082003 (2012)
[28] T.G.F. Li *et al.*, J. Phys. Conf. Ser. **363**, 012028 (2012)
[29] C. Van Den Broeck, *Probing dynamical spacetimes with gravitational waves*, to appear in Springer Handbook of Spacetime, eds. A. Ashtekar and V. Petkov, Springer Verlag, Berlin, 2013
[30] `https://www.lsc-group.phys.uwm.edu/daswg/projects/lalsuite.html`
[31] A. Buonanno, B.R. Iyer, E. Ochsner, Y. Pan, and B.S. Sathyaprakash, Phys. Rev. D **80**, 084043 (2009)
[32] Yi Pan *et al.*; arXiv:1307.6232 [gr-qc]
[33] M. Hannam *et al.*; arXiv:1308.3271 [gr-qc]
[34] R. Sturani *et al.*, J. Phys. Conf. Ser. **243**, 012007 (2010)
[35] R. Sturani *et al.*, arXiv:1012.5172 [gr-qc]
[36] K.S. Thorne, in *300 Years of Gravitation*, eds. S.W. Hawking and W. Israel, Cambridge University Press, Cambridge, England, 1987
[37] B.S. Sathyaprakash and S.V. Dhurandhar, Phys. Rev. D **44**, 3819 (1991)
[38] C. Van Den Broeck and A.S. Sengupta, Class. Quantum Grav. **24**, 155 (2007)
[39] C. Van Den Broeck and A.S. Sengupta, Class. Quantum Grav. **24**, 1089 (2007)
[40] R. O'Shaughness, B. Farr, E. Ochsner, H.-K. Cho, C. Kim, and C.-H. Lee; arXiv:1308.4704 [gr-qc]
[41] J. Veitch and A. Vecchio, Phys. Rev. D **78**, 022001 (2008)
[42] J. Veitch and A. Vecchio, Class. Quant. Grav. **25**, 184010 (2008)
[43] J. Veitch and A. Vecchio, Phys. Rev. D **81**, 062003 (2010)
[44] LIGO Scientific Collaboration, Virgo Collaboration, Astrophys. J. **760**, 12 (2012)
[45] LIGO Scientific Collaboration, Virgo Collaboration; Phys. Rev. D, **88** 6, 062001 (2013), arXiv:1304.1775 [gr-qc]
[46] D.R. Lorimer, Liv. Rev. Rel. **11**, 8 (2008)
[47] G. Faye, L. Blanchet, and A. Buonanno, Phys. Rev. D **74**, 10403 (2006)
[48] A. Bohé, S. Marsat, and L. Blanchet; arXiv:1303.7412
[49] K.G. Arun, A. Buonanno, G. Faye, and E. Ochsner, Phys. Rev. D **79**, 104023 (2009); Erratum ibid. D **84**, 049901 (2011)
[50] M. Agathos, W. Del Pozzo, T.G.F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, to appear in Proceedings of the 13th Marcel Grossmann Meeting; arXiv:1305.2963 [gr-qc]
[51] M. Hannam *et al.*, Astrophys. J. **766**, L14 (2013)
[52] R. Valentim, E. Rangel, and J.E. Horvath; arXiv:1101.4872 [astro-ph]
[53] B. Kiziltan, A. Kottas, and S.E. Thorsett; arXiv:1011.4291 [astro-ph]
[54] M. Dominik *et al.*, Astrophys. J. **759**, 52 (2012)
[55] F. Beauville *et al.*, Class. Quantum Grav. **22**, 4285 (2005)
[56] D.A. Brown (for the LIGO Scientific Collaboration), Class. Quantum Grav. **22**, S1097 (2005)
[57] S. Babak, R. Balasubramanian, D. Churches, T. Cokelaer, and B.S. Sathyaprakash, Class. Quantum Grav. **23**, 5477-5504 (2006)
[58] T. Cokelaer, Phys. Rev. D **76**, 102004 (2007)
[59] T. Cokelaer, Class. Quantum Grav. **24**, 6227 (2007)
[60] C. Van Den Broeck *et al.*, Phys. Rev. D **80**, 024009 (2009)
[61] D.A. Brown, I. Harry, A. Lundgren, and A.H. Nitz, Phys. Rev. D **86**, 084017 (2012)
[62] S. Babak *et al.*, Phys. Rev. D **87**, 024033 (2013)
[63] D.A. Brown, P. Kumar, and A.H. Nitz, Phys. Rev. D **87**, 082004 (2013)
[64] C. Cutler *et al.*, Phys. Rev. Lett. **70**, 2984 (1993)
[65] M. Maggiore, *Gravitational Waves. Volume 1: Theory and Experiments*, Oxford University Press, Oxford, 2008
[66] A. Kolmogorov, G. Inst. Ital. Attuari **4**, 83 (1933)
[67] N.V. Smirnov, Ann. Math. Stat. **19**, 279 (1948)
[68] T. Hinderer, B.D. Lackey, R.N. Lang, and J.S. Read, Phys. Rev. D **81**, 123016 (2010)
[69] J.M. Lattimer, Annu. Rev. Nucl. Part. Sci. **62**, 485 (2012)
[70] W. Del Pozzo, T.G.F. Li, M. Agathos, C. Van Den Broeck, and S. Vitale, Phys. Rev. Lett. **111**, 071101 (2013)
[71] J.S. Read *et al.*; arXiv:1306.4065
[72] P. Ajith *et al.*, Class. Quantum Grav. **29**, 124001 (2012)
[73] L.E. Simone, S.W. Leonard, E. Poisson, and C.M. Will, Class. Quantum Grav. **14**, 237 (1997)
[74] S. Vitale *et al.*, Phys. Rev. D **85**, 064034 (2012)
[75] W.G. Laarakkers and E. Poisson, Astrophys. J. **512**, 282

(1999)
[76] J. Vines, T. Hinderer, and É.É. Flanagan, Phys. Rev.D **83**, 084051 (2011)
[77] T. Damour, A. Nagar, and L. Villain, Phys. Rev. D **85**, 123007 (2012)
[78] K. Chatziioannou, A. Klein, N. Yunes, and N. Cornish; arXiv:1307.4418 [gr-qc]
[79] L. Santamaria *et al.*, Phys. Rev. D **82**, 064016 (2010)
[80] Here we are referring to both statistical and systematic errors; see [45] for examples.
[81] This means that in the simulations, we don't quite use the cut $\mathcal{M} < 1.3\,M_\odot$ proposed above, but our key results are unlikely to be affected by this choice.
[82] Damour, Nagar, and Villain have computed tidal contributions to 2.5PN [77], but these were not yet available when this work was started.
[83] An exception is the so-called TaylorEt waveform, which is considered pathological for this reason.
[84] It is also possible that the contrary is true, since it is not known whether the post-Newtonian expansion converges [73].